CALDERA: A Red-Blue Cyber Operations Automation Platform

Ron Alford, Dean Lawrence, Michael Kouremetis

The MITRE Corporation
{ralford, dlawrence, mkouremetis}@mitre.org

Abstract

Live evaluation of cybersecurity defenses, or red team engagements, can be costly, difficult to commission, and inconsistent in scope, detail, and results. This high overhead prevents many organizations from fully using them despite their benefits. CALDERA enables automated assessment of a network's susceptibility to an adversary being successful, essentially allowing an organization to see their network through the eyes of attackers on demand. CALDERA features an adversary model that maps to the MITRE ATT&CK® framework and an extensible planning system able to select and execute techniques. Inspired by automated planning methodologies, CALDERA provides a flexible, mature platform for developing adaptive and intelligent cyber agents.

Offensive cyber testing, often called "red team" or "adversary emulation," is a critical component of good cyber hygiene. In these exercises, testers ("red teamers") will attempt to attack a system to understand its weaknesses, informing the system's defenders and enabling them to harden the system before an actual attack. Unfortunately, red team testing is hard to execute in practice. Actual engagements can be costly, both from a monetary and a time perspective, and require highly trained staff to execute. Operational constraints make these engagements difficult to design and repeat.

Automation technologies can help ease the burden on organizations looking to deploy red teaming. In particular, automated red team or automated adversary emulation technologies can lower the barrier to entry to running offensive cyber tests by being more cost- and time-effective, removing staff from in-the-loop to on-the-loop, and allowing for quick and repeatable design and results analysis.

Industry has recognized this benefit, and a new field of tools has emerged looking at automating the red team process (Yoo et al. 2020). Oftentimes these tools contain scripted sequences of techniques, allowing organizations to replay specific scenarios, but not to stray beyond these baked-in operating paradigms. At the same time, researchers have recognized that automated planning can be leveraged to intelligently and dynamically compose attacks (Hoffmann 2015). These technologies can enable the same tool to execute in multiple ways, better stress testing a network and providing a more realistic adversary.

The open-source CALDERA[1] automated adversary emulation software is built around this idea, featuring many planning-inspired techniques built into the code. CALDERA integrates many capabilities needed for an automated red teaming suite, including a library of techniques to execute, a custom implant for execution, multiple command and control mechanisms, etc., alongside a custom-built planning module and API. CALDERA features a robust set of parsers and a custom logic that allows it to learn and leverage new information during an operation, effectively encoding requirements and consequences for many of its techniques. This latter planning feature makes CALDERA fairly unique, allowing it to not only compose knowledge gained during an operating into new attacks, but also for users to customize how exactly they want CALDERA to make decisions during an operation.

CALDERA Architecture

CALDERA is cross-platform framework for running cybersecurity operations. A central server provides an interface for both users and coordination for CALDERA agents distributed across the network. The server can manage multiple simultaneous operations, for both offensive (red) and defensive (blue) groups. Operations are defined by an adversary profile (a set of available abilities, or operators) and choice of planner, which instantiates abilities as sets of instructions for agents to execute. As it starts, each operation maintains its own set of facts and relationships between facts. When an agent checks in with the server, it collects instructions assigned by its operation's selected planners.

CALDERA agents are implants responsible for executing abilities on host machines. Agents are a dynamic part of an operation. A typical offensive operation might start with a single agent on an initial host. As an offensive operation progresses, new agents may be implanted on compromised hosts, and other agents may be killed by an opposing operation or machine failure. As soon as an agent starts, it begins checking in with the server for waiting instructions.

Abilities define the actions that agents may take, such as dumping usernames and passwords, mounting remote file shares, and scheduling remote tasks. Abilities are defined in YAML files containing a name, description, and a set of

[1] https://github.com/mitre/caldera/

platform-dependent instructions and requirements (preconditions). Instructions define a command string with variables (parameters) to be replaced, an optional payload (e.g., for executables), and a parser for ingesting the command output (the reported effects). The flexible definition of abilities allows new capabilities to be defined with relatively little code.

An operation's planner chooses the instructions that are assigned to each agent. Although previous versions of CALDERA have included goal-oriented planning (Applebaum et al. 2016; Miller et al. 2018), ability development has outpaced our ability to provide the full declarative models required by traditional planning techniques. Instead, the currently included planners are oriented around either executing all abilities in a profile until completion, either by looping through them by a fixed priority, or bucketing abilities into discrete stages to simulate phases of an attack.

Written in Python, the CALDERA server provides a plugin interface, through which much of its core functionality is implemented. All planners, abilities, and agents are distributed through plugins for CALDERA, and a documented tool and process exists for creating new plugins. Third parties have created new abilities, including one for deploying cryptominers, and new planning functionality, such as an overlay for a reinforcement learning framework (Li, Fayad, and Taylor 2021).

Related and Future Work

CALDERA is distributed with nearly a thousand distinct abilities, many of which are imported automatically from the Atomic Red Team library (Smith 2017). Although much of their function is defined in a declarative fashion, requirements (preconditions) and parsing (effects) reference python objects, making it difficult to automatically translate from their YAML definitions to PDDL. Learning full action models from the YAML ability templates and action traces would provide a solid basis for the application of goal-oriented planning and model-based reinforcement learning techniques. Declarative semantics for abilities would enable two additional lines of research: validating new abilities against current sets during development, and explicitly planning around the partially-observable, open environment of red teaming (Miller et al. 2018).

More advanced reasoning techniques are required as defenders evaluate the impact of deploying deceptive measures on their networks, such as honeypot servers (Provos 2004) and fake credentials (Herley and Florêncio 2008). Simple priority-based execution strategies will be easily thrown off by fake objects, repeatedly trying facts until exhaustion, which would be a poor analogue to human adversaries. Automated planners are more selective with the facts used, making effective deception strategies more difficult to predict (Alford and Applebaum 2021).

Multiple efforts are underway to evaluate the effect of defensive response and deception in simulated environments (Walter, Ferguson-Walter, and Ridley 2021; CAGE 2021). These simulations abstract away time, replacing it with the game concept of turns which may not exist in an environment with fast-paced actions and observation delay. CALDERA can provide a platform to evaluate the effectiveness of realtime defensive response on emulated production networks and software.

Conclusion

CALDERA is a mature platform for automating cyber operations, adopted in both academic research and industry (GitLab 2021). We hope CALDERA can provide strong link between the cybersecurity and planning communities, exposing the planning community to the unique challenges and scope of cyber environments, while providing scalable, robust decision making for autonomous cyber agents. We will be providing a tutorial for CALDERA at ICAPS, as well as an updated demo on using CALDERA to evaluate defensive cyber deception.

References

Alford, R.; and Applebaum, A. 2021. Towards Causal Models for Adversary Distractions. In *AI/ML for Cybersecurity: Challenges, Solutions, and Novel Ideas at SDM '21*.

Applebaum, A.; Miller, D.; Strom, B.; Korban, C.; and Wolf, R. 2016. Intelligent, automated red team emulation. In *Proc. of ACSAC*, 363–373.

CAGE. 2021. CAGE Challenge 1. In *IJCAI-21 1st International Workshop on Adaptive Cyber Defense*.

GitLab. 2021. Red Team: Building Operation Capabilities. https://about.gitlab.com/handbook/engineering/security/security-operations/red-team/.

Herley, C.; and Florêncio, D. 2008. Protecting financial institutions from brute-force attacks. In *IFIP International Information Security Conference*, 681–685.

Hoffmann, J. 2015. Simulated Penetration Testing: From" Dijkstra" to" Turing Test++". In *Proc. of ICAPS*.

Li, L.; Fayad, R.; and Taylor, A. 2021. CyGIL: A Cyber Gym for Training Autonomous Agents over Emulated Network Systems. In *IJCAI-21 1st International Workshop on Adaptive Cyber Defense*.

Miller, D.; Alford, R.; Applebaum, A.; Foster, H.; Little, C.; and Strom, B. 2018. Automated adversary emulation: A case for planning and acting with unknowns. In *ICAPS Workshop on Integrated Planning, Acting and Execution*.

Provos, N. 2004. A Virtual Honeypot Framework. In *USENIX Security Symposium*, volume 173, 1–14.

Smith, C. 2017. Red Canary Introduces Atomic Red Team, a New Testing Framework for Defenders. https://redcanary.com/blog/atomic-red-team-testing/.

Walter, E.; Ferguson-Walter, K.; and Ridley, A. 2021. Incorporating Deception into CyberBattleSim for Autonomous Defense. In *IJCAI-21 1st International Workshop on Adaptive Cyber Defense*.

Yoo, J. D.; Park, E.; Lee, G.; Ahn, M. K.; Kim, D.; Seo, S.; and Kim, H. K. 2020. Cyber Attack and Defense Emulation Agents. *Applied Sciences*, 10(6).